Bioinformatics, 36(16), 2020, 4383–4388
doi: 10.1093/bioinformatics/btaa548
Advance Access Publication Date: 29 May 2020
Discovery Note

OXFORD

Databases and ontologies

Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons

Xiaoqiong Wei [1,2,†], Chengxin Zhang [2,†], Peter L. Freddolino[2,3,*] and Yang Zhang[2,3,*]

[1]State Key Laboratory of Biotherapy and Cancer Center/Collaborative Innovation Center of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China, [2]Department of Computational Medicine and Bioinformatics and [3]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Zhiyong Lu

Received on September 3, 2019; revised on March 24, 2020; editorial decision on May 23, 2020; accepted on May 26, 2020

Abstract

Motivation: Many protein function databases are built on automated or semi-automated curations and can contain various annotation errors. The correction of such misannotations is critical to improving the accuracy and reliability of the databases.

Results: We proposed a new approach to detect potentially incorrect Gene Ontology (GO) annotations by comparing the ratio of annotation rates (RAR) for the same GO term across different taxonomic groups, where those with a relatively low RAR usually correspond to incorrect annotations. As an illustration, we applied the approach to 20 commonly studied species in two recent UniProt-GOA releases and identified 250 potential misannotations in the 2018-11-6 release, where only 25% of them were corrected in the 2019-6-3 release. Importantly, 56% of the misannotations are 'Inferred from Biological aspect of Ancestor (IBA)' which is in contradiction with previous observations that attributed misannotations mainly to 'Inferred from Sequence or structural Similarity (ISS)', probably reflecting an error source shift due to the new developments of function annotation databases. The results demonstrated a simple but efficient misannotation detection approach that is useful for large-scale comparative protein function studies.

Availability and implementation: https://zhanglab.ccmb.med.umich.edu/RAR.

Contact: petefred@umich.edu or zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Due to the rapid accumulation of protein sequences and the slow experimental characterization of their functions, the majority of proteins can only be annotated by computational analysis. As of UniProt-GOA (Huntley *et al.*, 2015) release 2019-06-03, for example 99% of the GO term annotations have an evidence code 'Inferred from Electronic Annotation (IEA)', which refers to the terms assigned by fully automated computational pipelines, such as InterPro protein family searching (Jones *et al.*, 2014).

An almost inevitable by-product from the utilization of computational function annotations is the misannotation of protein functions in large databases (Andorf *et al.*, 2007; Schnoes *et al.*, 2009). While misannotation is a generally acknowledged challenge, estimations of annotation error rates vary widely from study to study. This variability is in part due to the heterogeneity of the sources of GO term annotation. The GO consortium coordinates the GO term annotation efforts of UniProt (Bateman *et al.*, 2019) and 31 other contributing groups (http://geneontology.org/docs/annotation-contributors/), all with different standards and approaches of annotation. The resulting UniProt-GOA database mainly consists of three kinds of annotations distinguished by their evidence codes: expert-curated GO terms derived from experimental literature, which have evidence codes EXP, IDA, IMP, etc.; computationally derived GO terms that undergo expert review, which have evidence codes ISS, IBA, RCA, etc.; and fully computational GO terms, which have the evidence code IEA. Expert curated annotations obtained from experimental evidence have the highest quality: Swiss-Prot annotation is close to error-free (Schnoes *et al.*, 2009), while error rates of annotations derived from experimental literature are estimated to be 1.82 and 1.40% for the CGD and EcoCyc databases, respectively (Keseler *et al.*, 2014). On the other hand, electronic GO annotations, including both fully automatically predicted IEA terms and expert-reviewed computational terms, are considered less reliable.

For example, by investigating the consistency of GO terms among homologs, as annotated by the GO consortium in 2007-03-03, an early study (Jones *et al.*, 2007) claimed that 30% of GO annotations could be imprecise or erroneous, i.e. with misassigned terms which are parents, children or irrelevant to the correct GO terms, and that reviewed computational terms with 'Inferred from Sequence or structural Similarity' (ISS) evidence were more error prone than unreviewed IEA terms. Another survey (Skunca *et al.*, 2012) on 747 154 annotations from 2008-01-16 that were later removed on 2011-01-11 also concluded that reviewed computational terms are less reliable (i.e. more likely to be removed in a later release) than unreviewed terms. While these studies were performed a decade ago and may not reflect the current (greatly improved) quality of GO annotations, as shown in our later section, they nevertheless highlight the issues of GO misannotations, which have not been completely corrected in the interim.

To reduce misannotations, various approaches have been proposed by the GO consortium (Huntley *et al.*, 2014). For example, taxon-based constraints were proposed to detect inconsistency in function annotations, which resulted in the removal of many erroneous assignments (Deegan *et al.*, 2010). Although important, such taxonomic constraints cannot be comprehensive enough to detect the ubiquitous annotation errors, due to the substantial manual efforts required to create and enforce these taxon-based rules.

In this study, we proposed a ratio of annotation rate (RAR)-based approach to detect potential taxon-specific inconsistency in a large set of GO term annotations by automated comparison of annotation rate of GO terms across different taxa. For 20 commonly studied species, 250 potential misannotations were identified and manually confirmed by our approach. Notably, 140 (56%) of the potential misannotations have 'Inferred from Biological aspect of Ancestor' (IBA) evidence from semi-manual phylogenetic analysis (Gaudet *et al.*, 2011). Our findings highlight the need for more stringent taxon-specific function annotation consistency checking, especially those derived by phylogenetic analysis; we also provide a computational framework to perform an initial consistency screening with minimal human effort.

2 Materials and methods

An illustration of the general idea for RAR-based GO misannotation detection is outlined in Figure 1. First, we classify all protein GO term annotations by different taxon groups, such as the animal, bacteria, archaea, fungi and plant kingdoms (as demonstrated below, finer-grained taxonomic distinctions can also be used). For a GO term q, its annotation rate in taxon t can be calculated as

$$P_t(q) = n_t(q)/N_t \quad (1)$$

where N_t is the total number of annotated proteins in taxon t, and $n_t(q)$ is the subset of proteins annotated with q. If q is annotated to at least two taxa, its RAR can be calculated by

$$\mathrm{RAR}(q) = \min_t \{P_t(q)\} / \max_t \{P_t(q)\} \quad (2)$$

Here, a smaller RAR indicates a greater possibility of the GO term being misannotated to the low annotation rate taxon. For example since the presence of a nucleus is typical of fungi but not bacteria, GO:0005634 'nucleus' is rare in bacteria (annotated to 6 out of 19922 bacterial proteins; annotation rate 6/19922 = 3.01E-4), but is common in fungi (annotation rate 0.347). Thus, this GO term has a low RAR (3.01E-4/0.347 = 8.68E-4), and is likely to be misannotated to bacteria. This RAR is also statistically significant, with a P-value < 2.22E-16 by rate ratio test (Fay, 2010) (see Supplementary Text S1).

While a low RAR is suggestive of incorrect GO term assignment, manual confirmation of the potentially incorrect cases is often necessary. In fact, a superficially low RAR can either come from biases in curation where a function is rarely annotated to a taxon simply due to lack of comprehensive experimental literature on the taxon

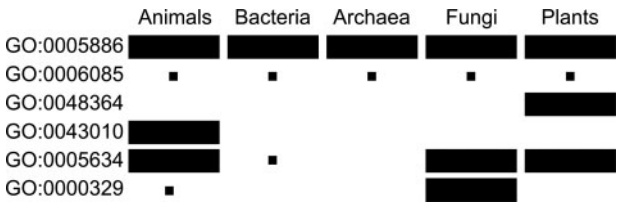

Fig. 1. Illustration of the taxon-specific, RAR-based GO term misannotation detection approach. The area of the rectangle is proportional to the number of proteins annotated with a GO term (row) in a taxon (column). The first two GO terms are unlikely to be inconsistently annotated because different taxa have similar the portion of annotated proteins. The next two GO terms are also disregarded by our analysis because each GO term is only annotated to one taxon. The last two GO terms will be picked up by the RAR analysis because it is common in at least one taxon but rare in another taxon

(Schnoes *et al.*, 2013), or from host-pathogen interactions that are easily overlooked. For example, GO:0061630 'ubiquitin protein ligase activity' is a rare GO term in bacteria because ubiquitin-dependent protein degradation is a eukaryote-specific protein catabolic pathway. The RAR for this GO term is low (6.54E-3) and significant (P-value =1.41E-05) in our dataset, as there is only one bacterial protein (SspH2, UniProt ID: P0CE12) with the GO term, compared to the 1014 for animals. However, the bacterial annotation is in fact correct in this case, because SspH2 is an E3 ubiquitin ligase, which interferes with ubiquitination pathways in eukaryotic host upon *Salmonella* infection (Quezada *et al.*, 2009). To avoid incorrectly flagging rare but correct terms, we manually inspected every GO term with RAR <0.1 to confirm whether the GO term is indeed a misannotation, as detailed below.

3 Results

3.1 Datasets

We studied potential misannotations in two recent UniProt-GOA releases 2019-06-03 (ftp.ebi.ac.uk/pub/databases/GO/goa/old/UNIPROT/goa_uniprot_all.gaf.189.gz) and 2018-11-06 (ftp.ebi.ac.uk/pub/databases/GO/goa/old/UNIPROT/goa_uniprot_all.gaf.183.gz). Here, we use two releases separated by half a year to investigate how many misannotations in the old release were corrected in a later release. Root terms (GO:0003674 'molecular_function', GO:0008150 'biological_process' and GO:0005575 'cellular_component'), the extremely common GO:0005515 'protein binding' and annotations with 'NOT' qualifiers are excluded because they either are too general or indicate lack of function. For this study, we focus on reference proteomes of 5, 7, 3, 3 and 2 species of animals, bacteria, archaea, fungi and plants, respectively (Supplementary Table S1), chosen from among the best studied model organisms and common pathogens in their respective taxa.

3.2 Overall statistics of potential misannotations

For each of the two UniProt-GOA releases, we performed a kingdom-level RAR analysis across five kingdoms (animals, bacteria, archaea, fungi and plants). Another phylum-level analysis is performed within the animal kingdom by rate ratio analysis between vertebrates (3 species) and invertebrates (2 species). Manual inspection of the 2731 and 2856 GO terms with RAR <0.1 (Supplementary Table S2) in kingdom- or phylum-level analysis confirmed 190 and 250 potentially misannotated in 2019-06-03 and 2018-11-06, respectively (Table 1, Supplementary Tables S3–S6). We henceforth use the term 'potential misannotation' to refer specifically to the human-confirmed subset of the initial annotations flagged by our pipeline. About 53 and 60% of the confirmed GO terms from the respective releases are significant in term of P-values (<0.05) after False Discovery Rate correction (Supplementary Text S1).

Since the subset of proteins used in our analysis come from a set of relatively commonly studied species, the portion of IEA GO terms

in our dataset (55%, Fig. 2A) is lower than that in the whole UniProt-GOA database (99%). Nevertheless, in our 20 reference proteomes, annotations with IEA evidence code constitute more than half of the proteins, while no other evidence code is associated with >11% of annotations (Fig. 2). Strikingly, our analysis shows that GO terms with IBA evidence are particularly susceptible to taxon inconsistency, despite a previous study attributing non-IEA misannotations mainly to ISS evidence code (Jones *et al.*, 2007). Such inconsistency is partly caused by difference in the use of IBA versus ISS evidence in different UniProt-GOA versions: release 2007-03-03 studied by Jones et al. (ftp.ebi.ac.uk/pub/databases/GO/goa/old/UNIPROT/gene_association.goa_uniprot.37.gz) did not have any IBA term; on the other hand, in release 2018-11-06 used in this study, 56% of all reviewed terms are IBA terms. Out of the 250 potential misannotations for release 2018-11-06, 140 annotations (56%) have IBA evidence code, surpassing in number the 93 (37%) IEA annotations. These data suggest that the sanity check performed by our taxon-specific RAR analysis is useful for filtering both IBA and IEA GO terms.

While most potential misannotations are acquired through computational modeling (such as IEA and IBA terms), GO terms from experimental literature curation occasionally also contain error. For example, human Junction Plakoglobin (JUP, UniProt ID: P14923) was assigned a non-animal GO term GO:0005199 'structural constituent of cell wall' with evidence code 'Inferred by Curator' (IC), based on a study on the role of JUP in cadherin/catenin complexes assembly (Sacco *et al.*, 1995). The curator assigning this GO term probably (incorrectly) associated the catenin complex, a cell surface protein complex, with cell wall, which was not implied by the original experimental article. Such cases of over-interpretation of literatures by curators are rare, but are still worthy of attention, and can be captured by our RAR-based analysis.

Of all 43 GO terms in the misannotations of release 2019-06-03, 12 terms violate existing taxon constraints curated by the GO consortium before the release date (https://github.com/geneontology/go-ontology/tree/master/src/taxon_constraints). Another nine violate the most recent taxon constraints release on 2020-03-11. This suggests that our RAR is complementary to existing taxon constraint curation efforts by the GO consortium, and that the taxon constraints are not universally enforced in all GO annotation processes.

Table 1. Overall statistics of potential misannotations identified by our RAR approach

UniProt-GOA release	Analysis type	Number of potential misannotations		
		GO terms	Proteins	Annotations[a]
2019-06-03	Kingdom	31	100	109
	Phylum	12	81	81
	Both	43	181	190
2018-11-06	Kingdom	37	153	170
	Phylum	13	80	80
	Both	50	233	250

[a]'Annotations' refers to the number of protein-GO term associations. For example, if GO:0005739 'mitochondrion' and GO:0005634 'nucleus' are both misannotated to two proteins P39615 and P12295, this table will count 2 GO terms, 2 proteins and 4 annotations.

3.3 Case studies on the cause of potential misannotations in phylogenetic analysis

As shown in the previous section, compared to other GO term types, IBA GO terms are more prone to potential misannotations. In this section, we explore the cause of potential misannotations specific to IBA GO terms, using Uracil-DNA glycosylase (Udg, UniProt ID: P39615) from *B.subtilis* as an example. Even though *B.subtilis* is a bacterium (prokaryote), its Udg protein is annotated with two IBA GO terms typical of eukaryotes: GO:0005739 'mitochondrion' and GO:0005634 'nucleus', both of which are misannotations. According to the 'WITH/FROM' field of UniProt-GOA, these two GO terms are assigned based on Uracil-DNA Glycosylate (Udg) protein family [PANTHER database ID: PTN000137400 (Mi *et al.*, 2019)] using the PAINT semi-manual phylogeny-based function annotation application (Gaudet *et al.*, 2011). The GO terms are ultimately derived from 6 orthologous proteins in the same family: RGD : 1307200, MGI : 109352, WB: WBGene00013241, PomBase: SPCC1183.06, SGD: S000004483 and TAIR : 2086904, which are from three animals (*R.norvegicus*, *M.musculus* and *C.elegans*), two fungi (*S.pombe*, and *S.cerevisiae*) and a plant (*A.thaliana*), respectively. To understand why eukaryotic proteins were used to annotate the prokaryote target, we check the phylogenetic tree of this family in PANTHER database (Fig. 3).

In the PANTHER phylogenetic tree, proteins from animals and fungi as well as a portion of the plant proteins are grouped to one branch consisting solely of eukaryotes (first three leaf nodes of Fig. 3A). Meanwhile, the remaining plant proteins (fourth node) are grouped into the branch of bacteria, and are located in the sub-branch of proteobacteria. Since experimentally annotated GO terms are usually propagated to the most recent common ancestor in the

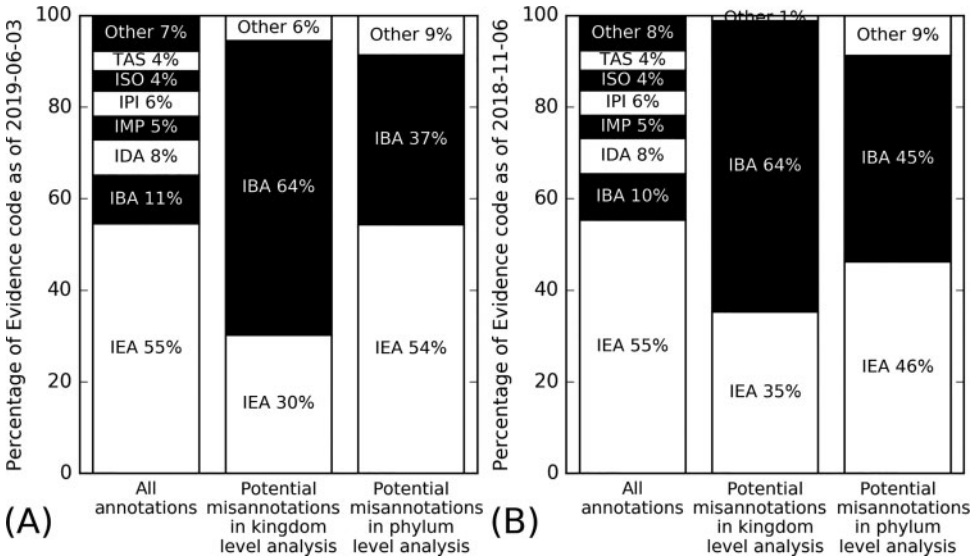

Fig. 2. Distribution of evidence codes in GO term annotations for the 20 species analyzed in this study in UniProt-GOA release (**A**) 2019-06-03. (**B**) 2018-11-06

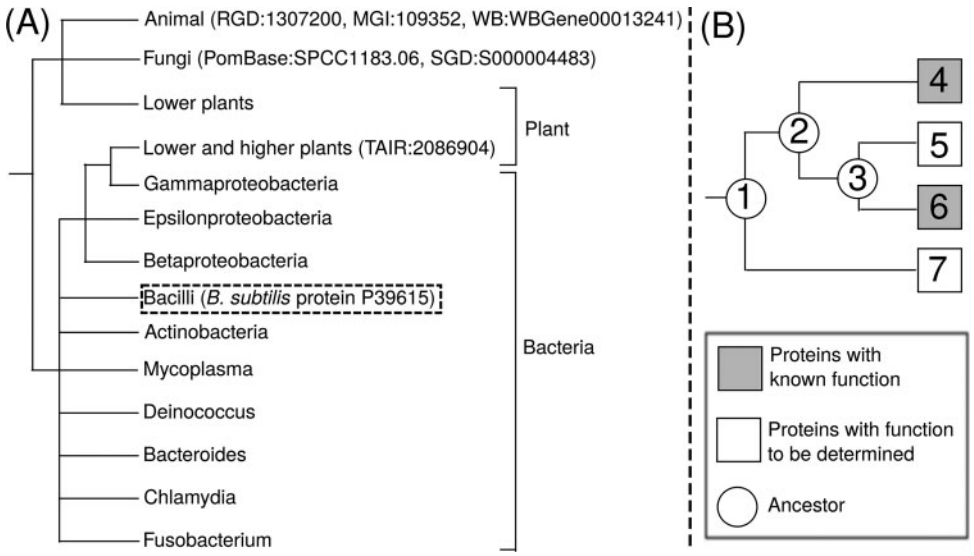

Fig. 3. Phylogeny based annotation of IBA GO terms. (**A**) A simplified phylogenetic tree for Udg protein family (PTN000137400 in PANTHER database). The six orthologous proteins with experimentally annotated functions are shown in parentheses. The *B.subtilis* target protein to be annotated with IBA GO term is shown in parentheses with dashed box. The full phylogenetic tree for this PANTHER family is provided as Supplementary Figure S1 to S5. (**B**) A diagram for GO term annotation using a phylogenetic tree for a protein family with four member proteins (Squares 4–7), where two proteins (Squares 4 and 6 in grey) have the same experimentally annotated GO terms, while the function of the other two proteins (Squares 5 and 7 in white) are to be determined. Among the inferred biological ancestors (Circle 1–3) in this phylogenetic tree, ancestor 2 is the most recent common ancestor (MRCA) of Proteins 4 and 6. Upon manual inspection of the tree, curators often infer that the whole branch rooted by MRCA (Ancestor 2) share the same GO term as the leaf proteins (Protein 4–6), and assign the GO term annotation of Proteins 4 and 6 to Protein 5. The function of Protein 7, which belongs to an outgroup and does not have the same MRCA as 4 and 6, is usually left unassigned

PAINT method (Fig. 3B), the 'mitochondrion' and 'nucleus' GO terms are propagated to the root node of the whole tree, and hence the whole UdG family is associated with both GO terms. As explained in Supplementary Figure S1 caption, the potential misannotations of prokaryotic proteins from the phylogenetic tree are probably not caused by incorrect phylogenetic tree construction, but by over-interpretation of its functional implication.

Misannotations of the 'mitochondrion' and 'nucleus' terms to bacteria affect not only this *B. subtilis* protein, but also its orthologs in *H.pylori* (UniProt ID: P56397), and *E.coli* (UniProt ID: P12295). A thorough search of the two terms through all species in UniProt-GOA revealed 453 and 466 potential misannotations in releases 2018-11-06 and 2019-06-03, respectively (Supplementary Table S7). Since these annotations have been in UniProt-GOA long enough (at least as early as 2017-02-28), they have been propagated to secondary databases such as BsubCyc (https://biocyc.org/gene?orgid=BSUB&id=BSU37970-MONOMER#tab=GO), *H.pylori* Pathway/Genome Database (https://helicobacter.biocyc.org/gene?orgid=HPY&id=HP1347-MONOMER#tab=GO) and EcoCyc (https://ecocyc.org/gene?orgid=ECOLI&id=EG11058-MONOMER#tab=GO), even after the correction of respective UniProt-GOA entry for the *E.coli* ortholog.

While the above case study mainly discussed misannotation of cellular component terms, phylogenetic analysis also affects annotation of molecular function and biological process. For example, the *A.thaliana* protein At3g08840 (UniProt ID: A0A1I9LPE3) is annotated with IBA terms GO:0008716 'D-alanine-D-alanine ligase activity' for molecular function and GO:0009252 'peptidoglycan biosynthetic process' for biological process. Both terms are assigned by PAINT method based on phylogenetic tree built for D-alanine-D-alanine ligase (DDL) family (PANTHER ID: PTN000566166), and the annotation is ultimately derived from DdlA and DdlB proteins of *E.coli* (UniProt IDs: P0A6J8 and P07862) and Ddl protein of *M.tuberculosis* (UniProt ID: P9WP31). These bacterial orthologs are used by the bacteria to ligate alanine residues in order to form peptidoglycan (Bruning *et al.*, 2011; Zawadzke *et al.*, 1991), which are building blocks of bacterial cell wall. However, in plants such as *A.thaliana*, the cell wall is made of cellulose instead of peptidoglycan, indicating the annotation of these two terms, especially the biological process term 'peptidoglycan biosynthetic process', to the plant ortholog At3g08840 is likely to be incorrect.

3.4 Correction of misannotations by UniProt and other members in the GO consortium

Function annotations are constantly subjected to correction. While some of these corrections could be caused by technical reasons such as changes in UniProt accession mapping and are not necessarily for removal of misannotations, they nevertheless reflect the extensive efforts by the Gene Ontology consortium (Huntley *et al.*, 2014). For the 20 species analyzed in this study, 4.5% of GO term annotations originally presented in UniProt-GOA releases 2018-11-06 were later corrected (i.e. removed) in 2019-06-03. Among corrected GO term annotations, 81 and 13% have IEA and IBA evidence codes, respectively, while each of the other remaining evidence codes are associated with <1% of the corrected annotations (Fig. 4A). 63 (25%) of the 250 potential misannotations flagged by our RAR analysis for release 2018-11-06 were corrected in 2019-06-03 (Fig. 4B), where 47 are IBA annotations. Moreover, 55 and 53% of the misannotations in the two releases have been in the UniProt-GOA database for more than 1 year (Supplementary Fig. S6). These data suggest that, despite extensive curatorial efforts for correcting misannotations, additional quality control measures, such as the RAR analysis presented here, are needed.

To assess the current scale of misannotation problems in common databases, we used the correction of UniProt-GOA annotations across the two releases to estimate the reliability of contemporary annotations. In UniProt-GOA 2018-11-06, 750 and 9773 annotations are rejected by 'NOT' qualifiers and confirmed by new low-throughput experiments, respectively, in UniProt-GOA 2019-06-03 (Supplementary Fig. S7). The error rate of annotations is thus estimated to be around 7% [=750/(750 + 9773)]; this is a likely an underestimate of the actual error rate, as negative experimental results are usually less likely to be published than positive results. Nevertheless, our data suggest significantly improved quality of computational GO annotation since a much higher error rate was reported more than a decade ago (Jones *et al.*, 2007). Among these 10 523 annotations, IBA terms have the highest estimated error rate (31%), followed by IEA terms (5%), while experimental annotations are almost error-free. This is consistent with our RAR analysis

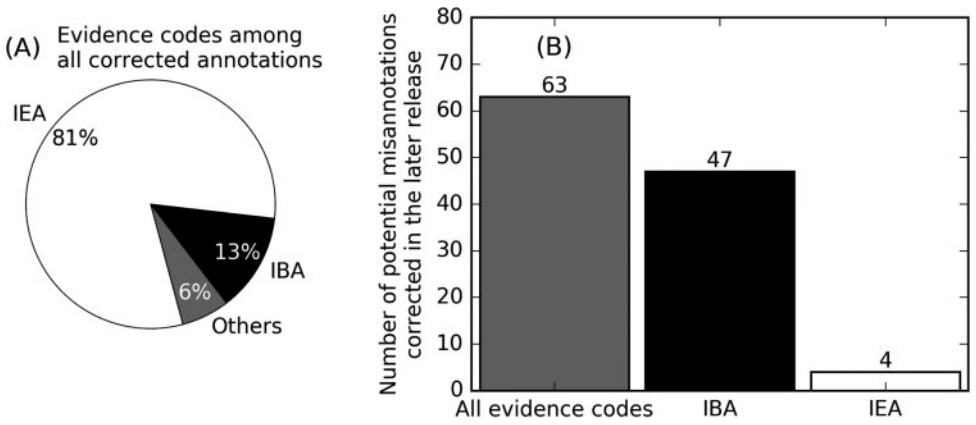

Fig. 4. Correction of GO terms in UniProt-GOA. (**A**) Distribution of evidence codes in GO terms in release 2018-11-06 that are removed in release 2019-06-03. (**B**) Number of GO term annotations removed in 2019-06-03 among potential misannotations in release 2018-11-06 flagged by our RAR analysis. Here, we do not include a GO term if it is not annotated to any proteins in any species of UniProt-GOA release 2019-06-03

as showed in Figure 2. Overall, the 250 misannotations identified by RAR with a high confidence counts only for a small fraction of the dataset (~0.01% = 250/1 984 375). Nevertheless, none of the 250 misannotations are among the 750 annotations rejected by 'NOT', indicating that the RAR analysis is highly complementary to existing curation efforts in UniProt and the GO consortium.

4 Conclusions

We developed a new pipeline for detecting potential misannotations in large sets of GO term annotations (such as UniProt-GOA). This method uses automated RAR analysis of taxon-specific GO term annotation, followed by manual inspection to identify function annotations that are not compatible with the organisms' taxon. Application of this pipeline on 20 commonly studied species in UniProt-GOA releases 2019-06-03 and 2018-11-06 reported 190 and 250 potential misannotations of GO terms, respectively. Among the potential misannotations flagged by our pipeline, the largest portion of annotations is manually curated IBA GO terms, followed by fully automatically annotated IEA GO terms. This contradicts an earlier study (Jones *et al.*, 2007) performed attributing the main source of annotation errors to ISS GO terms, and reflects the recent introduction of phylogeny-based function annotation (Gaudet *et al.*, 2011) for a substantial number of UniProt proteins. Our finding echoes a recent study (Skunca *et al.*, 2012), which concluded that, on average, annotations assigned by curators without experimental literature (e.g. IBA GO terms) are not more reliable than automated electronic annotations (i.e. IEA GO terms). One of the likely reasons of IBA term misannotations is the over-interpretation of phylogenetic trees during function curation, as shown by case studies on potential misannotations of prokaryotic GO terms to eukaryotic proteins in the Ddl family, and eukaryotic GO terms to prokaryotic proteins in the Udg family, where three misannotations have also spread through secondary protein function databases. A more recent UniProt-GOA release corrects only 25% of misannotations in the older release that we analyzed, and these corrections are not always reflected in secondary databases in a timely manner. The method developed herein can thus be used to systematically develop new taxon constraints, as the current taxon constraints only cover 1052 (2.4%) of all 44674 GO terms. It can also be used as an additional quality control step in large scale function prediction studies (Zhang *et al.*, 2018). We are working with UniProt, the GO consortium and neXtProt to correct these misannotations (Ignatchenko, A., Lane, L. personal communications). Following inspection of our data, neXtProt additionally decided to downgrade the status of IBA annotations in the future release of neXtProt from 'Gold' to 'Silver' to reflect its lower quality (Lane,L. personal communications). We will make future annual updates of the misannotation detection analysis described herein available to the GO consortium through our website at https://zhanglab.ccmb.med.umich.edu/RAR.

The number of potential misannotations identified by this study is much smaller than those reported previously (Jones *et al.*, 2007) (or what actually exist in the databases) due to two reasons. First, the previous study estimated the annotation error rate based on annotation inconsistencies without confirming which annotations are actually incorrect, while our RAR analysis focuses more on pinpointing the specific annotations that we can verify to be incorrect. This makes our approach more conservative in terms of asserting which GO terms are misannotated. Second, potential misannotations identified by this study are likely just the tip of the iceberg of all misannotations detectably by RAR, as we are currently only performing kingdom- and phylum-level analysis on a small set of well-studied species representing <0.7% of all annotations in UniProt-GOA. In fact, a simple check of 2 out of 50 potentially misannotated GO terms ('nucleus' and 'mitochondrion', Supplementary Table S7) across all proteins in UniProt-GOA reveals 453 potential misannotations, which are 19 times more misannotations than we identified for the same 2 terms among the 20 model organisms that we considered (Supplementary Table S3). In the future, we plan to extend our approach to the entire UniProt-GOA database based on different levels of taxon groups to obtain a systematic examination of the database.

Acknowledgements

The authors thank Dr. Lydie Lane and Dr. Gilbert S. Omenn for inspection of misannotated human proteins in the neXtProt database. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation [ACI1548562].

Funding

This work was supported by the National Institutes of Health [GM083107, GM136422 and AI134678 to Y.Z.], the National Science Foundation [DBI1564756 and IIS1901191 to Y.Z.] and the China Scholarship Council [201506240207 to X.W.]. The work was done when X.W. visited the University of Michigan.

Conflict of Interest: none declared.

References

Andorf,C. *et al.* (2007) Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics*, **8**, 284.

Bateman,A. *et al.* (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.

Bruning,J.B. *et al.* (2011) Structure of the Mycobacterium tuberculosis D-alanine: D-alanine ligase, a target of the antituberculosis drug D-cycloserine. *Antimicrob. Agents Chemother.*, **55**, 291–301.

Deegan,J.I. *et al*. (2010) Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics*, **11**, 530.

Fay,M.P. (2010) Two-sided exact tests and matching confidence intervals for discrete data. *R. J*., **2**, 53–58.

Gaudet,P. *et al*. (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Brief. Bioinf*., **12**, 449–462.

Huntley,R.P. *et al*. (2014) Understanding how and why the gene ontology and its annotations evolve: the GO within UniProt. *Gigascience*, **3**, 4.

Huntley,R.P. *et al*. (2015) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res*., **43**, D1057–D1063.

Jones,C.E. *et al*. (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, **8**, **170**.

Jones,P. *et al*. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Keseler,I.M. *et al*. (2014) Curation accuracy of model organism databases. *Database*, **2014**, bau058.

Mi,H.Y. *et al*. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*., **47**, D419–D426.

Quezada,C.M. *et al*. (2009) A family of Salmonella virulence factors functions as a distinct class of autoregulated E3 ubiquitin ligases. *Proc. Natl. Acad. Sci. USA*, **106**, 4864–4869.

Sacco,P.A. *et al*. (1995) Identification of Plakoglobin domains required for association with N-cadherin and alpha-catenin. *J. Biol. Chem*., **270**, 20201–20206.

Schnoes,A.M. *et al*. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol*., **5**, e1000605.

Schnoes,A.M. *et al*. (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput. Biol*., **9**, e1003063.

Skunca,N. *et al*. (2012) Quality of computationally inferred gene ontology annotations. *PLoS Comput. Biol*., **8**, e1002533.

Zawadzke,L.E. *et al*. (1991) Existence of two D-alanine-D-alanine ligases in *Escherichia coli*: cloning and sequencing of the DdlA gene and purification and characterization of the DdlA and DdlB enzymes. *Biochemistry*, **30**, 1673–1682.

Zhang,C. *et al*. (2018) Structure and protein interaction-based gene ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17. *J. Proteome Res*., **17**, 4186–4196.